

日本国特許庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日

Date of Application:

2002年12月27日

出願番号

Application Number:

特願2002-378956

[ ST.10/C ]:

[ JP2002-378956 ]

出願人

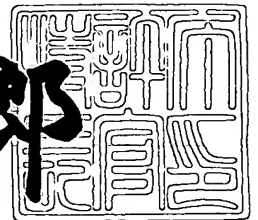
Applicant(s):

株式会社日立製作所

2003年 5月13日

特許庁長官  
Commissioner,  
Japan Patent Office

太田信一郎



出証番号 出証特2003-3035591



特 2 0 0 2 - 3 7 8 9 5 6

【書類名】 特許願

【整理番号】 H02016111A

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 15/16

【発明者】

    【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所 システム開発研究所内

    【氏名】 細谷 睦

【特許出願人】

    【識別番号】 000005108

    【氏名又は名称】 株式会社 日立製作所

【代理人】

    【識別番号】 100075096

    【弁理士】

    【氏名又は名称】 作田 康夫

    【電話番号】 03-3212-1111

【手数料の表示】

    【予納台帳番号】 013088

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

    【物件名】 図面 1

    【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 高可用ディスク制御装置とその障害処理方法及び高可用ディスクサブシステム

【特許請求の範囲】

【請求項 1】

ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部を有し、前記複数のホストインターフェース部の各々は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、前記複数のディスクインターフェース部の各々は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行するディスク制御装置であって、

前記複数のホストインターフェース部と前記キャッシュメモリ部、及び、前記複数のディスクインターフェース部と前記キャッシュメモリ部との間が 1 つ以上のスイッチで構成されるスイッチ網を介して接続され、前記ホストインターフェース部、前記ディスクインターフェース部及び前記キャッシュメモリ部は、前記スイッチ網内で一意に決まるローカル ID と該 ID を変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記 ID を用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキャッシュメモリ部は、該複数のキャッシュメモリ部における障害発生の有無を監視するための障害監視機構と前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を有することを特徴とするディスク制御装置。

【請求項 2】

前記パス制御機構は前記複数のキャッシュメモリ部における障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御することを特徴とする請求項 1 記載のディスク制御装置。

【請求項 3】

ホストコンピュータとのインターフェースを有する複数のホストインターフェ

ース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部と、ホストインターフェース部及びディスクインターフェース部と前記キャッシュメモリ部との間のデータ転送に関する制御情報を格納する複数のリソース管理部を有し、前記複数のホストインターフェース部の各々は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、前記複数のディスクインターフェース部の各々は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行するディスク制御装置であって、

前記複数のホストインターフェース部、前記複数のディスクインターフェース部が1つ以上のスイッチで構成されるスイッチ網を介して前記キャッシュメモリ部と接続され、前記複数のホストインターフェース部、前記複数のディスクインターフェース部が前記スイッチ網を介して前記リソース管理部に接続され、前記ホストインターフェース部、前記ディスクインターフェース部、前記キャッシュメモリ部及び前記リソース管理部は、前記スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記IDを用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキャッシュメモリ部と前記複数のリソース管理部は、その障害発生の有無を監視するための障害監視機構と前記スイッチ内の前記フォワーディングテーブルを制御するためのパス制御機構を有することを特徴とするディスク制御装置。

#### 【請求項4】

前記パス制御機構は前記キャッシュメモリ部または前記リソース管理部における障害発生時に障害部位を回避するように前記スイッチ内の前記フォワーディングテーブルを制御することを特徴とする請求項3記載のディスク制御装置。

#### 【請求項5】

前記キャッシュメモリ部における障害を監視するための前記障害監視機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部の中にあることを特徴とする請求項3または4記載のディスク制御装置。

【請求項 6】

前記障害監視機構が、前記ホストインターフェース部及び前記ディスクインターフェース部でのリードまたはライトとともに動作することを特徴とする請求項 1 乃至 5 のいずれかに記載のディスク制御装置。

【請求項 7】

前記キャッシュメモリ部における障害を処理するためのバス制御機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部に備わっていることを特徴とする請求項 3 乃至 6 のいずれかに記載のディスク制御装置。

【請求項 8】

前記障害監視機構により障害を検出した際、前記バス制御機構により前記障害部位の前記ローカルIDと前記障害部位の機能を引き継ぐ前記交換部位の前記ローカルIDとを交換し、該ローカルID交換に対応して前記スイッチ網内経路を切り換えるよう前記スイッチ内フォワーディングテーブルを変更することを特徴とする請求項 1 乃至 7 のいずれかに記載のディスク制御装置。

【請求項 9】

ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部を有し、各ホストインターフェース部は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、各ディスクインターフェース部は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、

前記複数のホストインターフェース部と前記キャッシュメモリ部、及び、前記複数のディスクインターフェース部と前記キャッシュメモリ部との間が 1 つ以上のスイッチで構成されるスイッチ網を介して接続され、前記ホストインターフェース部、前記ディスクインターフェース部及び前記キャッシュメモリ部は、前記スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記IDを用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキ

キャッシュメモリ部は、障害監視機構とパス制御機構を有するディスク制御装置における前記キャッシュメモリ部の障害処理方法であって、

前記障害監視機構が前記キャッシュメモリ部における障害発生の有無を確認するステップと、該障害監視機構が該障害発生を前記パス制御機構に通知するステップと、該パス制御機構が前記通知された障害情報を解析するステップと、該パス制御機構が障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するステップとを有することを特徴とする障害処理方法。

【請求項 1 0】

ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、記憶装置とのインターフェースを有する複数のディスクインターフェース部と、前記記憶装置に対しリードまたはライトされるデータを一時的に格納する複数のキャッシュメモリ部と、ホストインターフェース部及びディスクインターフェース部と前記キャッシュメモリ部との間のデータ転送に関する制御情報を格納する複数のリソース管理部を有し、前記複数のホストインターフェース部の各々は、前記ホストコンピュータとのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、前記複数のディスクインターフェース部の各々は、前記記憶装置とのインターフェースと前記キャッシュメモリ部との間のデータ転送を実行し、

前記複数のホストインターフェース部、前記複数のディスクインターフェース部が 1 つ以上のスイッチで構成されるスイッチ網を介して前記キャッシュメモリ部と接続され、前記複数のホストインターフェース部、前記複数のディスクインターフェース部が前記スイッチ網を介して前記リソース管理部に接続され、前記ホストインターフェース部、前記ディスクインターフェース部、前記キャッシュメモリ部及び前記リソース管理部は、前記スイッチ網内で一意に決まるローカル ID と該 ID を変更する変更手段を有し、前記スイッチは前記スイッチ網内での経路を前記 ID を用いて指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、さらに前記複数のキャッシュメモリ部と前記複数のリソース管理部は、障害監視機構とパス制御機構を有するディスク制御装置における前記キャッシュメモリ部および前記リソース管理部の障害処理方法であって、

前記障害監視機構が、前記キャッシュメモリ部または前記リソース管理部における障害発生の有無を確認するステップと、該障害監視機構が該障害発生を前記パス制御機構に通知するステップと、該パス制御機構が、前記通知された障害情報を解析するステップと、該パス制御機構が障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するステップを有することを特徴とする障害処理方法。

【請求項 1 1】

請求項 1 0 記載のディスク制御装置の前記キャッシュメモリ部における障害を監視するための前記障害監視機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部の中にあるディスク制御装置における障害処理方法であって、該障害監視機構が前記キャッシュメモリ部の障害発生の有無を確認するステップが、前記リソース管理部の中で実行されることを特徴とする障害処理方法。

【請求項 1 2】

前記障害監視機構による障害発生の有無を確認するステップが、前記ホストインターフェース部及び前記ディスクインターフェース部でのリードまたはライト動作の際に実行されることを特徴とする請求項 9 乃至 1 1 のいずれかに記載の障害処理方法。

【請求項 1 3】

請求項 1 0 乃至 1 2 のいずれかに記載のディスク制御装置の前記キャッシュメモリ部における障害を処理するためのパス制御機構が、前記キャッシュメモリ部の中にかえて、前記リソース管理部に備わっているディスク制御装置における障害処理方法であって、前記パス制御機構による前記フォワーディングテーブルを制御するステップが、前記リソース管理部の中で実行されることを特徴とする障害処理方法。

【請求項 1 4】

前記パス制御機構により前記フォワーディングテーブルを制御するステップは、前記障害部位のローカル ID と障害部位の機能を引き継ぐ交換部位のローカル ID とを交換するステップと、該ローカル ID 交換に対応して前記スイッチ網内経路を切り換えるよう前記スイッチ内フォワーディングテーブルを変更するステップを

有することを特徴とする請求項 9 乃至 1 3 のいずれかに記載の障害処理方法。

【請求項 1 5】

前記記憶装置は磁気ディスク装置であることを特徴とする請求項 1 乃至 8 のいずれかに記載のディスク制御装置。

【請求項 1 6】

前記記憶装置は磁気ディスク装置であることを特徴とする請求項 9 乃至 1 4 のいずれかに記載の障害処理方法。

【請求項 1 7】

複数のホストコンピュータと第 1 のネットワークを介して接続され、複数の磁気ディスク装置と第 2 のネットワークを介して接続されたディスク制御装置を備え、

上記ディスク制御装置は、

上記ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、

上記磁気ディスク装置とのインターフェースを有する複数のディスクインターフェース部と、

上記複数のホストインターフェース部及び上記複数のディスクインターフェース部との間が 1 つ以上のスイッチで構成されるスイッチ網を介して接続された複数のキャッシュメモリ部を備え、

前記複数のホストインターフェース部、前記複数のディスクインターフェース部及び前記複数のキャッシュメモリ部は、前記スイッチ網内で一意に決まるローカル I D と該 I D を変更する変更手段を有し、

前記スイッチは、前記スイッチ網内での経路を前記 I D を用いて指定するためのフォワーディングテーブルと、該テーブルを変更する変更手段を有し、

前記複数のキャッシュメモリ部は、

該複数のキャッシュメモリ部における障害発生の有無を監視するための障害監視機構と、

前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を有することを特徴とするディスクアレイサブシステム。

## 【発明の詳細な説明】

【 0 0 0 1 】

## 【発明の属する技術分野】

本発明は、データを複数の磁気ディスク装置に格納するディスクシステム装置の制御装置に関する。

【 0 0 0 2 】

## 【従来の技術】

企業間の電子商取引や社会基盤としての金融システムなどでは、高度な信頼性が要求され、その根幹をなす基幹ストレージ・システムに対しても、極めて高い可用性が求められている。これら基幹ストレージ・システムでは、その可用性を高めるために、内部を冗長構成とし、障害発生時には自動的に故障個所を切り離して、正常な冗長部分で動作を継続するための自動障害回復機能を備えたディスク制御装置が広く使われている。

【 0 0 0 3 】

例えば、図 9 に示した、従来より知られているディスク制御装置は、ホストコンピュータ 6 0 との間のデータ転送を実行する複数のホストインターフェース部 1 X と、磁気ディスク装置 7 0 との間のデータ転送を実行する複数のディスクインターフェース部 2 X と、磁気ディスク装置 7 0 のデータを一時的に格納するキャッシュメモリ部 3 X と、ディスク制御装置 1 0 4 に関する制御情報（例えば、ホストインターフェース部 1 X 及びディスクインターフェース部 2 X とキャッシュメモリ部 3 X との間のデータ転送制御に関する情報、磁気ディスク装置 7 0 に格納するデータの管理情報）を格納するリソース管理部 5 X とを備えている。

ホストインターフェース部 1 X 及びディスクインターフェース部 2 X とキャッシュメモリ部 3 X との間は、データインターフェース信号 6 により接続される。ホストインターフェース部 1 X とキャッシュメモリ部 3 X との間、及び、ディスクインターフェース部 2 X とキャッシュメモリ部 3 X との間の接続にスイッチ 4 X を用いることもある。ホストインターフェース部 1 X 及びディスクインターフェース部 2 X とリソース管理部 5 X との間は、管理インターフェース信号 7 により接続される。リソース管理部 5 X と、ホストインターフェース部 1 X、及び、ディスク

インターフェース部 2 X との接続は、スイッチを介しても、介さなくても良い。これにより、リソース管理部 5 X およびキャッシュメモリ部 3 X は全てのホストインターフェース部 1 X 及びディスクインターフェース部 2 X からアクセス可能な構成となる。

#### 【 0 0 0 4 】

図 1 2 に示すように、ホストインターフェース部 1 X は、ホストインターフェース信号 1 との入出力を処理するチャンネルプロトコル処理部 9 0、データインターフェース信号 6 との入出力を処理するための内部プロトコル処理部 8 X、管理インターフェース信号 7 との入出力を処理するためのプロセッサインターフェース 1 7、及びホストコンピュータ 6 0 に対する入出力を制御するプロセッサ 1 4 とローカルメモリ 1 5 を有している。

ディスクインターフェース部 2 X も、構造としては、ホストインターフェース部と同様であるが、ホストインターフェース信号 1 の代わりにディスクインターフェース信号 2 がチャンネルプロトコル処理部 9 0 に接続され、プロセッサ 1 4 においては、ホストインターフェース部で行われる制御に加えて、RAID 機能の実行も行う。

ホストインターフェース部 1 X、及び、ディスクインターフェース部 2 X が、キャッシュメモリ部 3 X と通信を行う場合、データの先頭に宛先アドレスを付加したパケットを使用してパケット転送を行う。

ホストインターフェース部 1 X、もしくは、ディスクインターフェース部 2 X 内のプロセッサ 1 4 の制御で生成されたパケットは、データインターフェース信号 6 を介して、スイッチ 4 X に送られる。スイッチ 4 X は、図 1 0 で示したように、データインターフェース信号 6 に接続した複数のバスインターフェース 4 1 X と、パケットバッファ 4 3 とアドレスラッチ 4 4 とセレクト 4 8 を備えている。バスインターフェース 4 1 X 内には、パケットからアドレス情報を取り出すヘッダ解析部 4 2 X が含まれており、それによって解析抽出されたパケットアドレスがアドレスラッチ 4 4 に取り込まれる。一方、送られてきたパケットは、バスインターフェース 4 1 X を通して、パケットバッファ 4 3 に格納される。アドレスラッチ 4 4 からは、パケット宛先に応じたセレクト制御信号 4 7 が生成され、パケ

ットバッファ 4 3 に格納されたパケットの送出先をセクタ 4 8 によって切り換える。

#### 【 0 0 0 5 】

スイッチ 4 X で、宛先ごとに振り分けられたパケットは、再び、データインターフェース信号 6 を介して、目的のキャッシュメモリ部 3 X に転送される。キャッシュメモリ部 3 X は、図 1 1 で示したように、データインターフェース信号 6 に接続した複数のデータバスインターフェース 3 1 X と、パケットバッファ 3 3 と調停回路 3 9 とセクタ 3 8 を備えている。データバスインターフェース 3 1 X 内には、パケットからアドレス情報を取り出すヘッダ解析部 3 2 X が含まれており、それによって解析抽出されたパケットアドレスは、調停回路 3 9 に取り込まれる。一方、送られてきたパケットは、バスインターフェース 3 1 X を通して、パケットバッファ 3 3 に格納される。調停回路 3 9 は、複数のデータバスインターフェース 3 1 X の中から、いずれかを選択し、その選択結果に応じたセクタ制御信号を生成する。このセクタ制御信号で、セクタ 3 8 を切り換えることにより、所望のパケットバッファ 3 3 の内容をキャッシュメモリ 3 7 に、メモリ制御回路 3 5 を介して、書き込むことができる。パケットバッファ 3 3 に格納されたパケットがメモリ読み出しの要求であった場合、指定された領域のキャッシュメモリ 3 7 の内容を、上記と逆の経路をたどることによって、ホストインターフェース部 1 X、もしくは、ディスクインターフェース部 2 X に返送する。

#### 【 0 0 0 6 】

ホストインターフェース部 1 X、及び、ディスクインターフェース部 2 X が、リソース管理部 5 X と通信を行う場合も、データインターフェース信号 6 の代わりに、管理インターフェース信号 7 が用いられる以外は、キャッシュメモリ部との通信と同様なパケット転送が行われる。リソース管理部 5 X は、図 1 1 に示したキャッシュメモリ部と、インターフェース信号を除いて同等な構成になっている。

ここで、キャッシュメモリ部 3 X、及び、リソース管理部 5 X は、複数のホストインターフェース部 1 X、ディスクインターフェース部 2 X からアクセスされるシステム共通のリソースであり、その可用性がシステムの信頼性に大きな影響を与え

るため、同等機能を複数備えた冗長構成となっており、たとえ、一方に障害が発生しても、残りの正常な部分で動作を継続できるように設計されている。具体的には、ホストインターフェース部 1 X、もしくは、ディスクインターフェース部 2 X内のいずれかのプロセッサ 1 4 が、複数あるキャッシュメモリ部 3 X、もしくは、リソース管理部 5 Xのいずれかの障害を検知した場合、障害を検知したプロセッサ制御で、障害部分を閉塞して残りのキャッシュメモリ部 3 X、もしくは、リソース管理部 5 Xに、その機能を引き継がせるとともに、他の全てのプロセッサ 1 4 に対して障害通知を行う。障害通知を受けた全てのプロセッサは、障害に伴うシステム構成・通信経路の変更処理をそれぞれ行うことによって、すべてのホストインターフェース部 1 Xとディスクインターフェース部 2 Xで、障害部分の切り離しを実現することが可能となる。

このように従来のディスク制御装置 1 0 4 では、キャッシュメモリ部 3 X、もしくは、リソース管理部 5 Xのような共通リソース障害に伴うシステム構成・通信経路の変更処理が、複数あるホストインターフェース 1 X、および、ディスクインターフェース 2 X内のプロセッサそれぞれで分散して行われていた。そのため、共通リソースの障害においては、分散配置されたプロセッサ間でのブロードキャスト通信を含む複雑な処理が必要であった。

#### 【 0 0 0 7 】

ディスク制御装置を高信頼化する別の従来例として、共有システムリソースとシステムリソース・クライアントとの間で可用性の高いネットワーク通信を提供する障害処理機構が提案されている（例えば、特許文献 1 参照。）。この従来例においても、さきに説明した従来例と同様、複数あるプロセッサごとに経路変更（ルーティングテーブル変更）を行う方式となっている。

#### 【 0 0 0 8 】

また、ディスク制御装置の高可用性を実現する別の従来例として、ホストコンピュータとディスクアレイサブセットとの間に配置して、両者の間のアドレス変換を行うスイッチを備えた記憶装置システムが提案されている（例えば、特許文献 2 参照。）。この従来例では、複数あるディスクアレイサブセットに障害が発生した場合、スイッチ内でパケットの解釈を行って障害部分への要求を同等の機

能を有した冗長部分の宛先に変更することで、経路変更などの障害処理を行う方式となっている。

【特許文献 1】

特開 2 0 0 2 - 4 1 3 4 8 号公報

【特許文献 2】

特開 2 0 0 0 - 2 4 2 4 3 4 号公報

【0 0 0 9】

【発明が解決しようとする課題】

キャッシュメモリ部、リソース管理部のような共通リソースの障害は、ストレージ・システム、ひいては、ホストコンピュータで実行されているアプリケーションの動作不良を引き起こすため、すみやかな回復処理が行われなければならない。しかし、図 9、図 1 0、図 1 1、図 1 2 に示した従来技術では、すべてのホストインターフェース部 1 X、ディスクインターフェース部 2 X において経路変更が必要なため、障害処理に時間がかかり、ホストコンピュータとの間でのリード／ライトタスクの継続が行えずに、ストレージシステムの性能劣化やアプリケーション・プログラムの動作不良を引き起こす場合があった。また、この障害処理には、ホストインターフェース部 1 X、ディスクインターフェース部 2 X に高機能プロセッサと複雑な制御プログラムが必要となり、製造コストの増大、信頼性の低下を招いていた。特許文献 1 に示した別の従来技術においても、複数あるプロセッサごとにルーティングテーブル変更を行う必要があり、同様の課題があった。

【0 0 1 0】

また、特許文献 2 で開示されている別の従来技術では、パケット宛先変更機能を有するスイッチを導入することによって、複数あるディスクアレイ・サブセット間で機能を引き継がせる等の障害回避動作をスイッチ内処理のみで行うことが可能となる。しかし、その一方で、パケットごとに宛先の解釈が必要となり、障害発生時のみならず、通常動作時においても、処理に多くの時間を要し、ストレージシステムの性能劣化を招くという課題を有している。

【0 0 1 1】

本発明の目的は、上記従来技術の欠点を改善し、障害発生時に、迅速かつ高信頼に障害処理を行い、かつ、通常動作時においても性能劣化を起こさない高可用のディスク制御装置、および、その障害処理方法を提供することにある。

より具体的には、本発明の目的は、システム共通リソースの障害時を含め、いかなる場合においても、ストレージ・システムの性能劣化や、ホスト・アプリケーションの動作不良を引き起こすことのない高可用性ディスク制御装置を提供することにある。

#### 【 0 0 1 2 】

##### 【課題を解決するための手段】

上記目的を達成するため、本発明では、ホストコンピュータとのインターフェースを有する複数のホストインターフェース部と、磁気ディスク装置とのインターフェースを有する複数のディスクインターフェース部と、磁気ディスク装置に対しリード／ライトされるデータを一時的に格納するキャッシュメモリ部と、ホストインターフェース部及びディスクインターフェース部とキャッシュメモリ部との間のデータ転送に関する制御情報を格納するリソース管理部を有し、各ホストインターフェース部は、ホストコンピュータとのインターフェースとキャッシュメモリ部との間のデータ転送を実行し、各ディスクインターフェース部は、磁気ディスク装置とのインターフェースとキャッシュメモリ部との間のデータ転送を実行するディスク制御装置であって、

複数のホストインターフェース部とキャッシュメモリ部及び複数のディスクインターフェース部とキャッシュメモリ部との間が1つ以上のスイッチで構成されるスイッチ網を介して接続され、スイッチがスイッチ網内での経路を指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を有し、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部が、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を有し、さらに複数のキャッシュメモリ部が、その障害発生の有無を相互に監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を有するディスク制御装置を提供する。

## 【 0 0 1 3 】

## 【発明の実施の形態】

以下、大容量のデータの記憶装置として磁気ディスク装置を例にとって説明するが、大容量記憶装置として磁気ディスクに限られるものではなく、例えばDVDのような大容量記憶装置であって良い。

本発明の実施の形態の1つとして、好ましくは、前記複数のホストインターフェース部と前記キャッシュメモリ部との間、及び前記複数のディスクインターフェース部と前記キャッシュメモリ部との間を、1つ以上のスイッチで構成されるスイッチ網を介して接続し、スイッチにスイッチ網内での経路を指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を設け、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部に、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を設け、さらに複数のキャッシュメモリ部に、その障害発生の有無を監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を設ける。

また、好ましくは、前記複数のホストインターフェース部、前記複数のディスクインターフェース部、前記キャッシュメモリ部、前記リソース管理部との間を1つ以上のスイッチで構成されるスイッチ網を介して接続し、スイッチにスイッチ網内での経路を指定するためのフォワーディングテーブルと該テーブルを変更する変更手段を設け、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部に、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を設け、さらに複数のキャッシュメモリ部に、その障害発生の有無を監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォワーディングテーブルを制御するためのパス制御機構を設ける。

また、好ましくは、前記リソース管理部に、前記キャッシュメモリ部、もしくは、前記リソース管理部の障害を監視するための障害監視機構を設ける。

## 【 0 0 1 4 】

また、好ましくは、前記ホストインターフェース部及び前記ディスクインター

フェース部に前記キャッシュメモリ部、もしくは、前記リソース管理部の障害監視機構に対して障害を報告する機能を設ける。

また、好ましくは、前記リソース管理部に、前記キャッシュメモリ部の障害を処理するためのバス制御機構を設ける。

その他、本願が開示する課題、及びその解決方法は、発明の実施形態の欄及び図面により明らかにされる。

#### 【 0 0 1 5 】

以下、本発明の実施例を図面を用いて説明する。

#### 《実施例 1》

図 1、図 2、図 6、図 7、及び図 8 に、本発明の一実施例を示す。

図 2 に示したディスク制御装置 1 0 0 は、ホストコンピュータ 6 0 とのインターフェース部（ホストインターフェース部） 1 0 と、磁気ディスク装置 7 0 とのインターフェース部（ディスクインターフェース部） 2 0 と、キャッシュメモリ部 3 0、スイッチ 4 0、リソース管理部 5 0 を有し、ホストインターフェース部 1 0 及びディスクインターフェース部 2 0 と、キャッシュメモリ部 3 0 及びリソース管理部 5 0 との間は、スイッチ 4 0 を介して、内部インターフェース信号 4 で接続されている。すなわち、内部インターフェース信号 4 を介して、全てのホストインターフェース部 1 0、全てのディスクインターフェース部 2 0 から、全てのキャッシュメモリ部 3 0、あるいはリソース管理部 5 0 へアクセス可能な構成となっている。

#### 【 0 0 1 6 】

図 8 に示すように、ホストインターフェース部 1 0 は、ホストインターフェース信号 1 との入出力を処理するチャンネルプロトコル処理部 9 0、データインターフェース信号との入出力を処理するための内部プロトコル処理部 8 0 を有し、ホストコンピュータ 6 0 とキャッシュメモリ部 3 0 間のデータの転送、及びリソース管理部 5 0 との間の制御情報の転送を実行する。

ディスクインターフェース部 2 0 も、構造としては、ホストインターフェース部と同様であるが、ホストインターフェース信号 1 の代わりにディスクインターフェース信号 2 を有し、磁気ディスク装置 7 0 とキャッシュメモリ部 3 0 間のデ

ータの転送、及びリソース管理部 5 0 との間の制御情報の転送を実行する。

キャッシュメモリ部 3 0 は、図 7 に示すように、内部インターフェース信号 4 との入出力処理を行う内部プロトコル処理部 8 0 とプロセッサ 3 6 とキャッシュメモリ 3 7 とメモリ制御回路 3 5 と DMA エンジン 3 4 を有し、磁気ディスク装置 7 0 へ記録するデータや磁気ディスク装置から読み出したデータを一時的に格納する。

#### 【 0 0 1 7 】

リソース管理部 5 0 も、キャッシュメモリ部 3 0 と同等の構成を有し、システム構成などの管理制御情報を維持する。

#### 【 0 0 1 8 】

スイッチ 4 0 は、図 6 に示すように、内部インターフェース信号 4 に接続した複数のバスインターフェース 4 1 と、パケットバッファ 4 3 とアドレスラッチ 4 4 とセレクタ 4 8 を備え、ホストインターフェース部 1 0 およびディスクインターフェース部 2 0 と、キャッシュメモリ部 3 0 およびリソース管理部 4 0 との間の経路接続を行う。

なお、可用性向上のために、ホストインタフェース部 1 0、ディスクインターフェース部 2 0、キャッシュメモリ部 3 0、リソース管理部 5 0 を、それぞれ複数のポートを有する構成とし、スイッチ 4 0 との間にそれぞれ複数の転送経路を設けることも出来る。

#### 【 0 0 1 9 】

ホストインターフェース部 1 0、ディスクインターフェース部 2 0、キャッシュメモリ部 3 0、リソース管理部 5 0 の内部プロトコル処理部 8 0 には、内部インターフェース信号 4 の接続しているスイッチ網内での宛先を一意に特定するためのローカル I D ( L I D ) を保持する L I D 情報 8 1 が備わっている。

#### 【 0 0 2 0 】

一方、スイッチ 4 0 内には、ポート番号（バスインターフェース 4 1 の位置）と L I D との対応関係を示すフォワーディング・テーブル 4 6 が備わっている。フォワーディング・テーブル 4 6 の例を、図 1 ( 1 ) に示す。この例では、2 つのホストインターフェース 1 0 と 2 つのディスクインターフェース 2 0 が 2 つの

スイッチ 4 0 A、4 0 B を介して、2 つのキャッシュメモリ（共通リソース）3 0 A、3 0 B に接続している。ホストインターフェース 1 0、ディスクインターフェース 2 0、キャッシュメモリ 3 0 A、3 0 B はそれぞれ 2 つずつの内部インターフェース信号とそれに対応したローカル I D（L I D）情報を持っている。スイッチ 4 0 A、4 0 B は、それぞれ 8 つのポート（パスインターフェース 4 1）とそれに対応したポート番号を持っている。フォーワーディング・テーブル 4 6 とは、この L I D とポート番号との対応表であって、例えば、スイッチ 4 6 A のフォーワーディング・テーブル A では、L I D ①、③、⑤、⑦、⑨、(11) が、それぞれポート a、b、c、d、e、f と接続していることを示している。このフォーワーディング・テーブルを参照することにより、パケット宛先（L I D）によって、どのポートに送出すればよいかがわかる。

#### 【 0 0 2 1 】

内部インターフェース信号の接続しているスイッチ網は、例えば、キャッシュメモリ部 3 0 内のプロセッサ 3 6 で実行されるネットワーク管理プログラムによって、維持管理されている。ネットワーク内の L I D 情報 8 1 やスイッチ内のフォーワーディング・テーブル 4 6 は、内部インターフェース信号 4 を介して、ネットワーク管理プログラムにより、設定更新される。

本発明のディスク制御装置における通常動作の一例として、図 2、図 6、図 7、図 8 を用い、ホストコンピュータ 6 0 からディスク制御装置 1 0 0 を介して磁気ディスク装置 7 0 へ読み出し要求を発行する場合の動作について説明する。

まず、ホストコンピュータ 6 0 は、自身が接続されているホストインターフェース部 1 0 にデータの読み出し要求を発行する。要求を受けたホストインターフェース部 1 0 は、リソース管理部 5 0 にアクセスし、要求されたデータがどの磁気ディスク装置 7 0 内に格納されており、当該磁気ディスクがどのキャッシュメモリ部 3 0 に制御されているかを調べる。リソース管理部 5 0 には、要求データのアドレスからこれらの情報を検索するためのテーブルが格納されており、要求されたデータをもとに管轄のキャッシュメモリ部を調べることができる。次に、要求を受けたホストインターフェース部 1 0 では、要求データを管理しているキャッシュメモリ部 3 0 に、読み出し要求を転送する。キャッシュメモリ部 3 0 は

、キャッシュメモリ 37 に要求されたデータが格納されているかどうかを確認する。キャッシュメモリ部 30 内にデータが存在しなかった場合、プロセッサ 36 は、要求データを磁気ディスク装置 70 から読出し、キャッシュメモリ 37 に格納する。キャッシュメモリ部 30 は、キャッシュメモリ 37 内に格納された要求データを、ホストインターフェース部 10 まで転送し、ホストコンピュータ 60 に送る。

ホストインターフェース部 10 及びディスクインターフェース部 20 がキャッシュメモリ部 30、及びリソース管理部 50 とスイッチ 40 を介して通信する場合、パケット宛先に L I D 情報を用い、スイッチでの経路変更にはフォワーディング・テーブル 46 を使用する。

例えば、ホストコンピュータ 60 からの読み出し要求を、スイッチ 40 を介して、キャッシュメモリ部 30 に転送する場合、ホストインターフェース部 10 では、ホストインターフェース信号 1 を、チャネルプロトコル処理部 90 と内部プロトコル処理部 80 を介して、内部インターフェース信号 4 に変換する。その際、読み出し要求パケットの宛先には、送り先のキャッシュメモリ部の L I D を設定する。ホストインターフェース部からの要求パケットを受けたスイッチ 40 では、経路変更手段 45 内のフォワーディング・テーブル 46 に従い、ヘッダ解析部 42 で抽出されたパケット宛先に応じたセクタ制御信号 47 を生成し、パケットバッファ 43 に格納されたパケットの送出先をセクタ 48 によって切り換えて、所望のキャッシュメモリ部 30 に向けてパケットを転送する（図16）。

#### 【 0 0 2 2 】

ディスクインターフェース部 20 が、キャッシュメモリ部と通信を行う場合、もしくは、ホストインターフェース部 10、及び、ディスクインターフェース部 20 が、リソース管理部 50 と通信を行う場合も、ホストインターフェース部とキャッシュメモリ部との間の通信と同様なパケット転送が行われる。

なお、リソース管理部 50 と、ホストインターフェース部 10 およびディスクインターフェース部 20 との接続に使われているスイッチ 40 を含んだスイッチ網は、キャッシュメモリ部 30 とホストインターフェース部 10 およびディスクインターフェース部 20 との接続に使われているスイッチ網と同じであっても、別

の専用ネットワークであっても良い（図17）。また、スイッチ40を使わずに、直接接続する形態であっても構わない。

次に、本発明のディスク制御装置の特徴である障害回復動作の一例として、図1、図2を用い、2つのキャッシュメモリ部30間での障害監視機構とパス制御機構の動作について説明する。

キャッシュメモリ部30は、その可用性向上のため、同等機能を有するマスタとスレーブの2つのキャッシュメモリ部を有している。スレーブキャッシュメモリは、マスタキャッシュメモリに障害が発生した際に、その機能を引き継ぐためのもので、ホットスタンバイで動作している。これらマスタ・キャッシュメモリ部とスレーブ・キャッシュメモリ部は、スイッチ40を介して、相互にその動作を確認するための障害監視機構Cを有している。すなわち、一定間隔ごとに自分の動作状況を報告するパケットを生成し、お互い相手の動作状態を監視することができる。その動作の概要を図13に示す。障害監視機構では通信が行われる度に、正常なシーケンスとACKがチェックされ、異常が発見された場合には、主経路もしくは副経路を介して、パス制御機構に障害通知を行う。図2に示した構成では、マスタおよびスレーブキャッシュメモリ部が、それぞれパス制御機構を有しているので、そのいずれかのキャッシュメモリ部に異常が発生した場合、その障害情報は、ただちにもう一方のキャッシュメモリ部で検知することができる。障害を検知したキャッシュメモリ部では、障害を起こしているキャッシュメモリ部を閉塞するとともに、内部に備えたパス制御機構Pによって、システム構成を変更し、ホストインターフェース部10およびディスクインターフェース部20が、障害キャッシュメモリ部にアクセスしないように制御する。

以下、パス制御機構Pの仕組みを、図14を使って説明する。障害監視機構から障害通知を受けたパス制御機構は、その通知の妥当性を確認した後、可用性向上のため複数個設けてあるパス制御機構間で、障害情報の同期化を行う。その後、障害解析を行い、現時点で障害特定が可能かどうか判断する。特定が出来ない場合、特定可能になるまで障害処理を遅延させる。特定可能になった場合、障害通知情報によりアクセス障害なのか機能障害なのかを判断し、機能障害の場合、障

害部位を含んだ冗長部位間で、処理途中の J O B などの同期化処理を試みる。その後、障害部位に対するアクセス・パスを冗長部位に振り替える交替パス処理を行う。図 1 で具体的に説明すると、マスタキャッシュメモリ部で機能障害が発生しスレーブキャッシュメモリ部にフェールオーバーする場合は、以下ようになる。障害のない正常動作状態における、ホストインターフェース部 1 0、ディスクインターフェース部 2 0、キャッシュメモリ部 3 0 A、3 0 B、スイッチ 4 0 A、4 0 B の L I D とフォワーディング・テーブルの値は、図 1 ( 1 ) に示した通りである。ここで、マスタ・キャッシュメモリ部 3 0 A に障害が発生し、スレーブ・キャッシュメモリ部 3 0 B の障害監視機構によって、当該障害が検出された場合、3 0 B のパス制御機構 P により、3 0 A の機能を 3 0 B が引き継ぐとともに、3 0 A 宛のパケットを 3 0 B に振り替える制御を行う。すなわち、3 0 A の 2 つの L I D ( ⑨、( 10 ) ) と 3 0 B の 2 つの L I D ( ( 11 )、( 12 ) ) とを交換し、それに対応して、フォワーディング・テーブル 4 6 A と 4 6 B を変更する。その結果、L I D とフォワーディングテーブルは図 1 ( 2 ) のようになり、3 0 A へのアクセスは、すべて 3 0 B に振り向けられて、障害部位 3 0 A のシステムからの切り離しが完了する。3 0 A の動作を 3 0 B が引き継ぐには、3 0 A と 3 0 B の内容が一致している必要があるが、それらは、通常の同期動作により実現される。すなわち、常に 3 0 A と 3 0 B が同じ内容になるよう、両方に対して同じアクセス要求を発生させる方法や、定期的に両方でデータをコピーする方法が考えられる。

なお、リソース管理部 5 0 も同様の障害監視機構 C とパス制御機構 P を備え、同様の手順で障害回復動作を行うことができる。これら障害監視機構やパス制御機構は、キャッシュメモリ部 3 0、もしくは、リソース管理部 5 0 内のプロセッサ 3 6 によって実行される制御プログラムとして実現することができる。また、本実施例では、リソース管理部 5 0 を独立に設けた構成としているが、リソース管理部をキャッシュメモリ部 3 0 の中に設けることもできる。また、図 1 で、スレーブキャッシュメモリ部内の L I D とマスタキャッシュメモリ部内の L I D を交換するのではなく、スレーブ側でマスタ側の L I D を追加して持つようにすることも可能である。その場合、障害前のスレーブ側 L I D が障害後も有効になると

いうメリットがある。

【 0 0 2 3 】

本実施例によれば、キャッシュメモリ部 3 0 もしくはリソース管理部 5 0 の障害に対して、スイッチ 4 0 のフォワーディング・テーブルとキャッシュメモリ部 3 0 もしくはリソース管理部 5 0 の L I D の変更のみで、障害部位の切り離しが完了し、従来技術のように、複数のホストインターフェース部 1 0 及びディスクインターフェース部 2 0 の間でのブロードキャスト通信や複雑な制御は不要である。そのため、障害発生時においても、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがない。

また、本実施例のスイッチ内フォワーディング・テーブルは、障害発生時のみ変更されものであって、従来技術のように、通信を行うたびにパケットの宛先の解釈や変更が行われる複雑なスイッチを必要としない。そのため、障害のない通常の動作における性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

【 0 0 2 4 】

《実施例 2》

図 3 に、本発明の他の実施例を示す。

図 3 に示した実施例は、キャッシュメモリ部 3 0、及び、リソース管理部 5 0 が、それぞれ障害通知のための専用線であるハートビート信号 3 を備えていることと、内部インターフェース信号のスイッチ網が多段のスイッチ 4 0 で構成されていることを除いて、実施例 1 の図 2 に示す構成と同様である。ただし、キャッシュメモリ部 3 0 は、障害監視機構 C のみを備え、パス制御機構は実装されていない。リソース管理部 5 0 は、障害監視機構 C、パス制御機構 P とともに備えている。

キャッシュメモリ部、および、リソース管理部は、マスタとスレーブにより冗長構成を形成しており、基本的には同じデータ内容を保持している。ただし、キャッシュメモリ部におけるディスクからの読み出しデータについては、マスタとスレーブ間で同一の内容を保持していなくても構わない。

障害が発生した場合の処理は、基本的に、実施例 1 と同様であるが、その概要を図 1 8 で簡単に説明する。マスタ／スレーブ キャッシュメモリ部、および、マスタ／スレーブリソース管理部は、障害監視機構により、相互に障害発生の有無について定期的な確認を行っている。障害監視機構によって発見された障害は、リソース管理部のバス制御機構に通知される。バス制御機構では、通知された障害情報の解析を行い、障害部位の特定を行う。バス制御機構では、障害部位が特定できれば、スイッチ内にあるフォワーディングテーブルを制御することで、障害部位を回避するようにバスの設定を行うことで、障害部位の切り離しが完了する。

本実施例では、専用のハートビート信号 3 を設けることにより、相互にその動作を確認するための障害監視機構 C を、実施例 1 より単純な構成で実現することができる。すなわち、ハートビート信号 3 を通して、お互いの動作状況を、直接、監視することができる。そのため、マスタ、もしくは、スレーブのいずれかのキャッシュメモリ部、もしくはリソース管理部に異常が発生した場合、その障害情報は、より迅速にもう一方のキャッシュメモリ部、もしくは、リソース管理部で検知することができる。

また、本実施例では、キャッシュメモリ部 3 0 内で検出された障害情報が、スイッチ 4 0 を介して(マスタ)リソース管理部 5 0 のバス制御機構 P に通知され、リソース管理部内のバス制御機構 P によってキャッシュメモリ部 3 0 の障害回復処理が行われる。これにより、障害情報をリソース管理部 5 0 に集めて、より適切な障害回復処理を行うことが可能となる。

また、本実施例では、ホストインターフェース側とディスクインターフェース側でスイッチを分離することにより、ホスト側とディスク側で柔軟な接続数変更が可能になるとともに、より大規模な構成への対応が可能となる。

#### 【 0 0 2 5 】

本実施例によれば、前述の実施例と同様、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがなく、また、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが

可能である。

【 0 0 2 6 】

### 《実施例 3》

図 4 に、本発明の他の実施例を示す。

図 4 に示した実施例は、キャッシュメモリ部 3 0、及び、リソース管理部 5 0 が、ハートビート信号 3 を備えていないことと、内部インターフェース信号のスイッチ網が冗長構成になっていることを除いて、実施例 2 の図 3 に示す構成と同様である。ただし、キャッシュメモリ部 3 0 は、障害監視機構 C とパス制御機構を実装していない。リソース管理部 5 0 は、障害監視機構 C、パス制御機構 P ともに備えている。

本実施例では、キャッシュメモリ部 3 0 の障害監視も、リソース管理部 5 0 内の障害監視機構 C を用いて行う。その実現方法としては、リソース管理部の障害監視機構 C が、定期的にキャッシュメモリ部 3 0 へアクセスを行い、キャッシュメモリ部の動作状況を監視する方法や、ホストインターフェース部 1 0 およびディスクインターフェース部 2 0 からキャッシュメモリ部 3 0 へアクセスした際に障害が検出された場合、その障害情報をリソース管理部に報告する方法などが考えられる。また、本実施例では、各ホストインターフェース部、各ディスクインターフェース部に複数のポートを設け、スイッチも 2 重化することにより、ホストインターフェース部、および、ディスクインターフェース部とキャッシュメモリ部、もしくは、リソース管理部との間に複数のパスを設けてある。

これにより、リソース管理部、キャッシュメモリ部の機能障害だけでなく、リソース管理部、もしくは、キャッシュメモリ部とホストインターフェース部およびディスクインターフェース部との間のパス障害についても、障害回復を行えることになり、可用性をさらに高めることができる。

また、リソース管理部 5 0 に、障害監視機構とパス制御機構を集めることにより、より障害状況の的確な分析が行えることになり、適切かつ高信頼の障害回復処理が可能となる。

【 0 0 2 7 】

本実施例によれば、前述の実施例と同様、迅速、かつ、高信頼な障害回復処理

を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがなく、また、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

【 0 0 2 8 】

《実施例 4》

図 5 に、本発明の他の実施例を示す。

図 5 に示した実施例は、キャッシュメモリ部 3 0、及び、リソース管理部 5 0 が、ハートビート信号 3 を備えていないことと、複数のディスク制御サブユニット 2 0 0 を備えることを除いて、実施例 2 の図 3 に示す構成と同様である。複数のディスク制御サブユニット内のキャッシュメモリ部それぞれが、障害監視機構 C を備えている。リソース管理部 5 0 は、障害監視機構 C、パス制御機構 P とともに備えている。

本実施例では、ディスク制御サブユニット 2 0 0 ごとに、キャッシュを分散配置することで、キャッシュの使用効率（ヒット率）を高めて性能を向上するとともに、ホスト側とディスク側で柔軟にシステム規模を拡張することが可能となり、より高スケーラブルなシステムの提供が可能となる。

また、本実施例では、実施例 2 と同様、キャッシュメモリ部 3 0 の障害に伴う障害回復処理も、リソース管理部 5 0 内のパス制御機構 P を用いて行う。実施例 2、3 と同様、リソース管理部 5 0 に、障害情報を集めることにより、より障害状況の的確な分析が行えることになり、ディスク制御サブユニット 2 0 0 の個数を増加させた、より大規模なディスク制御装置においても、適切かつ高信頼な障害回復処理が可能となる。

【 0 0 2 9 】

本実施例によれば、前述の実施例と同様、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがなく、また、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

## 【 0 0 3 0 】

## 《実施例 5》

図 1 5 に、本発明の他の実施例を示す。

図 1 5 に示した実施例では、実施例 1 - 4 で記載のディスク制御装置が、複数のホストコンピュータとホストコンピュータ間ネットワークを介して接続し、複数の磁気ディスク装置と磁気ディスク装置間ネットワークを介して接続している。ホストコンピュータ間ネットワークには、ファイルシステムの処理を行うサーバ 1 1 0 (NASヘッド)、複数のディスク制御装置が管轄するストレージをまとめて管理するためのサーバ 1 2 0 (ディスク制御装置間仮想化エンジン)、データベース・インターフェース処理を行うためのサーバ 1 3 0 (DB機能付加エンジン)などが接続されることもある。これら、NASヘッド、仮想化エンジン、DB機能付加エンジンは、ディスク制御装置の中に実装しても構わない。

## 【 0 0 3 1 】

本実施例によれば、迅速、かつ、高信頼な障害回復処理を行えるディスク制御装置を用いることにより、極めて可用性が高く、性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことのない、ストレージシステムの提供が可能となる。

## 【 0 0 3 2 】

## 【発明の効果】

以上説明したように、本発明では、キャッシュメモリ部 3 0 もしくはリソース管理部 5 0 の障害に対して、スイッチ 4 0 のフォワーディング・テーブルとキャッシュメモリ部 3 0 もしくはリソース管理部 5 0 の L I D の変更のみで、障害部位の切り離しが完了し、従来技術のように、複数のホストインターフェース部 1 0 及びディスクインターフェース部 2 0 の間でのブロードキャスト通信や複雑な制御は不要である。そのため、障害発生時においても、迅速、かつ、高信頼な障害回復処理を実現することができ、ストレージシステムの性能劣化や、ホストコンピュータ上のアプリケーション動作不良を引き起こすことがない。

また、本発明において、スイッチ内フォワーディング・テーブルは、障害発生時にのみ変更されものであって、従来技術のように、通信を行うたびにパケット

の宛先の解釈や変更が行われる複雑なスイッチを必要としない。そのため、障害のない通常の動作においての性能劣化は皆無で、かつ、低コストで高信頼に製造することが可能である。

【 0 0 3 3 】

本発明では、障害監視機構による障害発生通知を、パス制御機構で解析し、フォワーディングテーブルを制御するので、柔軟なシステム構成に対応可能である。とくに、複数のディスク制御サブユニットが存在する大規模なディスク制御装置の場合でも、複数の障害監視機構からの障害情報をパス制御機構に集めることにより、より障害状況の的確な分析が行え、高信頼な障害回復処理が可能となる

【図面の簡単な説明】

【図 1】

本発明によるディスク制御装置における障害回復処理の動作原理を示す図である。

【図 2】

本発明によるディスク制御装置の構成を示す図である。

【図 3】

本発明によるディスク制御装置の構成を示す図である。

【図 4】

本発明によるディスク制御装置の構成を示す図である。

【図 5】

本発明によるディスク制御装置の構成を示す図である。

【図 6】

本発明によるディスク制御装置内のスイッチの構成を示す図である。

【図 7】

本発明によるディスク制御装置内のキャッシュメモリ部の構成を示す図である。

【図 8】

本発明によるディスク制御装置内のホストインターフェース部の構成を示す図である。

【図 9】

従来のディスク制御装置の構成を示す図である。

【図 1 0】

従来のディスク制御装置内のスイッチの構成を示す図である。

【図 1 1】

従来のディスク制御装置内のキャッシュメモリ部の構成を示す図である。

【図 1 2】

従来のディスク制御装置内のホストインターフェース部の構成を示す図である。

【図 1 3】

本発明の障害監視機構の動作を示す図である。

【図 1 4】

本発明のバス制御機構の動作を示す図である。

【図 1 5】

本発明のディスク制御装置を用いたストレージシステムの例を示す図である。

【図 1 6】

本発明のディスク制御装置においてホストコンピュータからキャッシュメモリ部へのコマンド送信の概要を示す図である。

【図 1 7】

本発明のディスク制御装置の構成を示す図である。

【図 1 8】

本発明のディスク制御装置の障害処理の概要を示す図である。

【符号の説明】

- 1 ホストインターフェース信号
- 2 ディスクインターフェース信号
- 3、5 ハートビート信号
- 4 内部インターフェース信号
- 6 データインターフェース信号
- 7 管理インターフェース信号

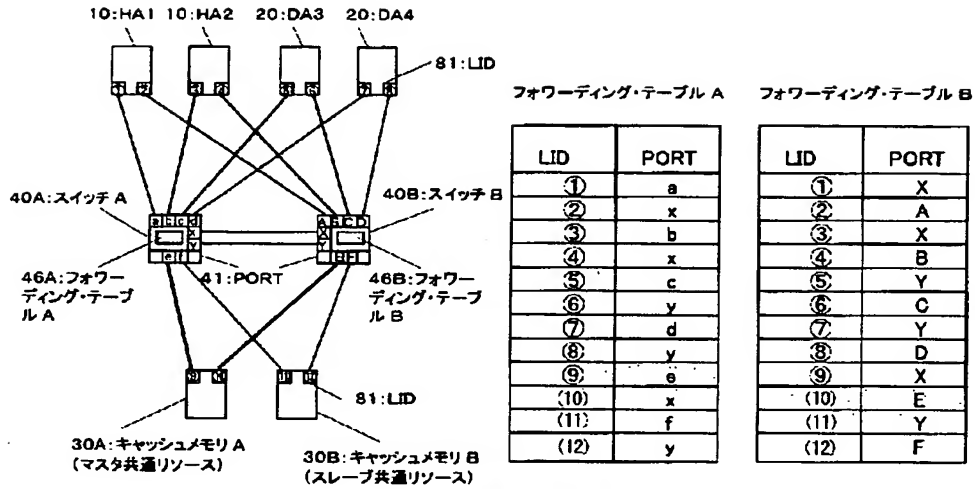
- 1 0、1 X ホストインターフェース部
- 1 1 DMAエンジン
- 1 2 送信データインターフェース
- 1 3 受信データインターフェース
- 1 4 プロトコル制御プロセッサ
- 1 5 ローカルメモリ
- 1 6 バスインターフェース
- 1 7 プロセッサインターフェース
- 2 0、2 X ディスクインターフェース部
- 3 0、3 0 A、3 0 B、3 X キャッシュメモリ部
- 3 1 X データバスインターフェース
- 3 2 X ヘッダ解析部
- 3 3 パケットバッファ
- 3 4 DMAエンジン
- 3 5 メモリ制御回路
- 3 6 プロセッサ
- 3 7 キャッシュメモリ
- 3 8 セレクタ
- 3 9 調停回路
- 4 0、4 0 A、4 0 B、4 X スイッチ
- 4 1 バスインターフェース (PORT)
- 4 1 X データバスインターフェース
- 4 2、4 2 X ヘッダ解析部
- 4 3 パケットバッファ
- 4 4 アドレスラッチ
- 4 5 経路変更手段
- 4 6、4 6 A、4 6 B フォワーディング・テーブル
- 4 7 セレクタ制御信号
- 4 8 セレクタ

- 5 0 リソース管理部
- 6 0 ホストコンピュータ
- 7 0 磁気ディスク装置
- 8 0、8 X 内部プロトコル処理部
- 8 1 ローカル I D 情報 ( L I D )
- 8 2 送信 P H Y
- 8 3 受信 P H Y
- 8 4 プロトコル制御プロセッサ
- 8 5 ローカルメモリ
- 8 6 バッファ
- 8 7 リンク処理
- 8 8 ヘッダ制御
- 9 0 チャネルプロトコル処理部
- 9 2 送信 P H Y
- 9 3 受信 P H Y
- 9 6 バッファ
- 9 7 リンク処理
- 9 8 トランSPORT処理
- 1 0 0 ~ 1 0 4 ディスク制御装置
- 1 1 0 N A S ヘッド
- 1 2 0 ディスク制御装置間仮想化エンジン
- 1 3 0 D B 機能付加エンジン
- 2 0 0 ディスク制御サブユニット
- C 障害監視機構
- P パス制御機構。

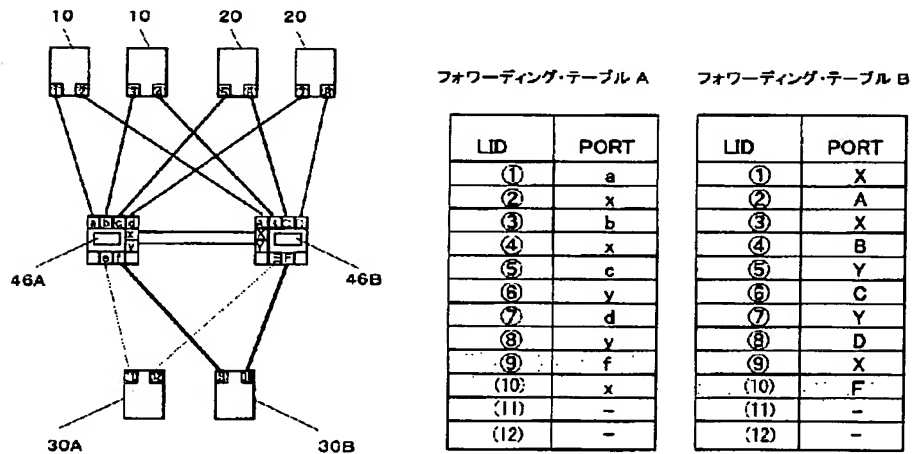
【書類名】 図面

【図 1】

図 1



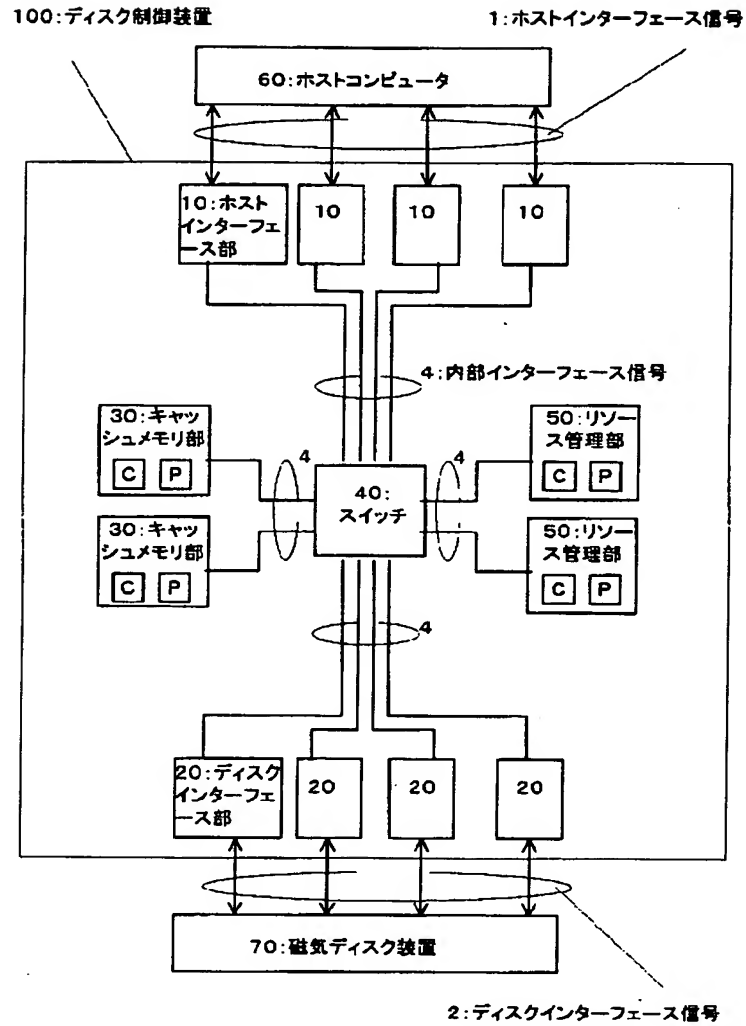
(1)障害前の状態



(2)30Aの障害処理後の状態

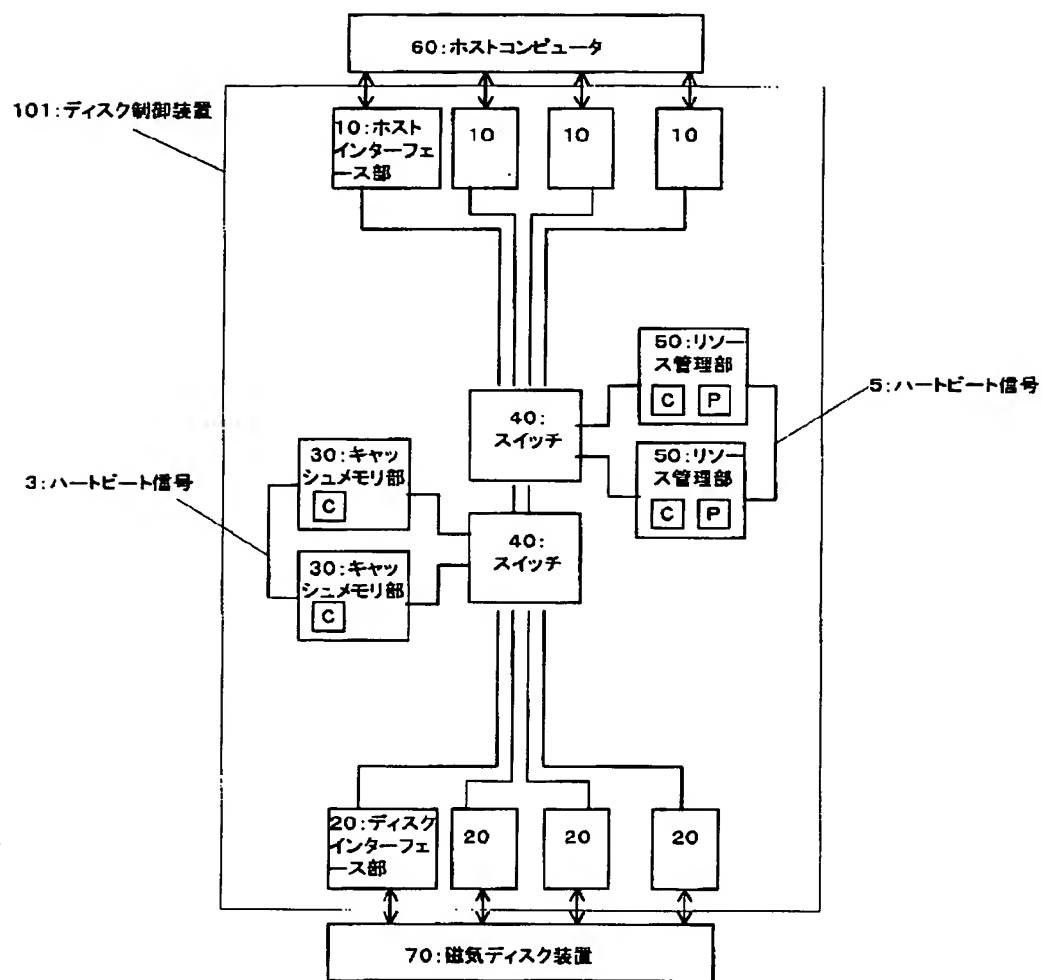
【図 2】

図2



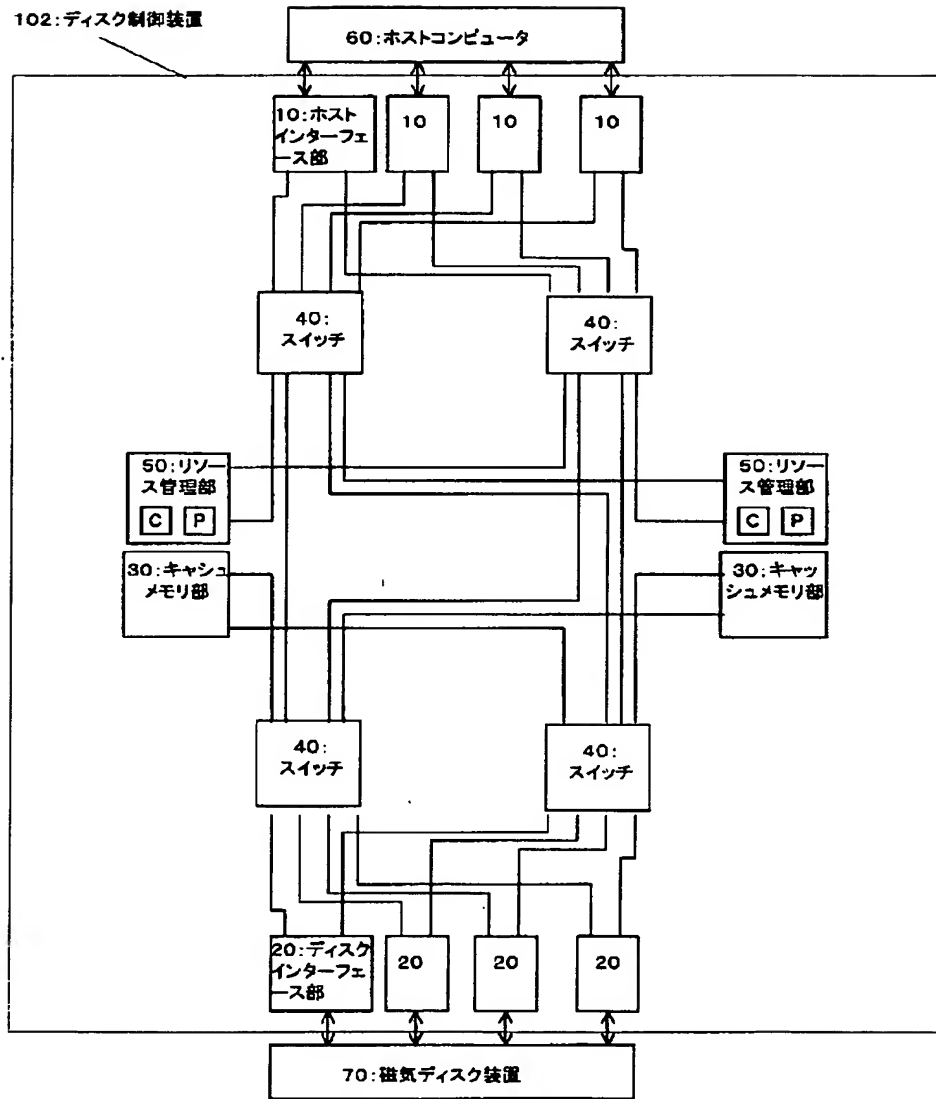
【図 3】

図3



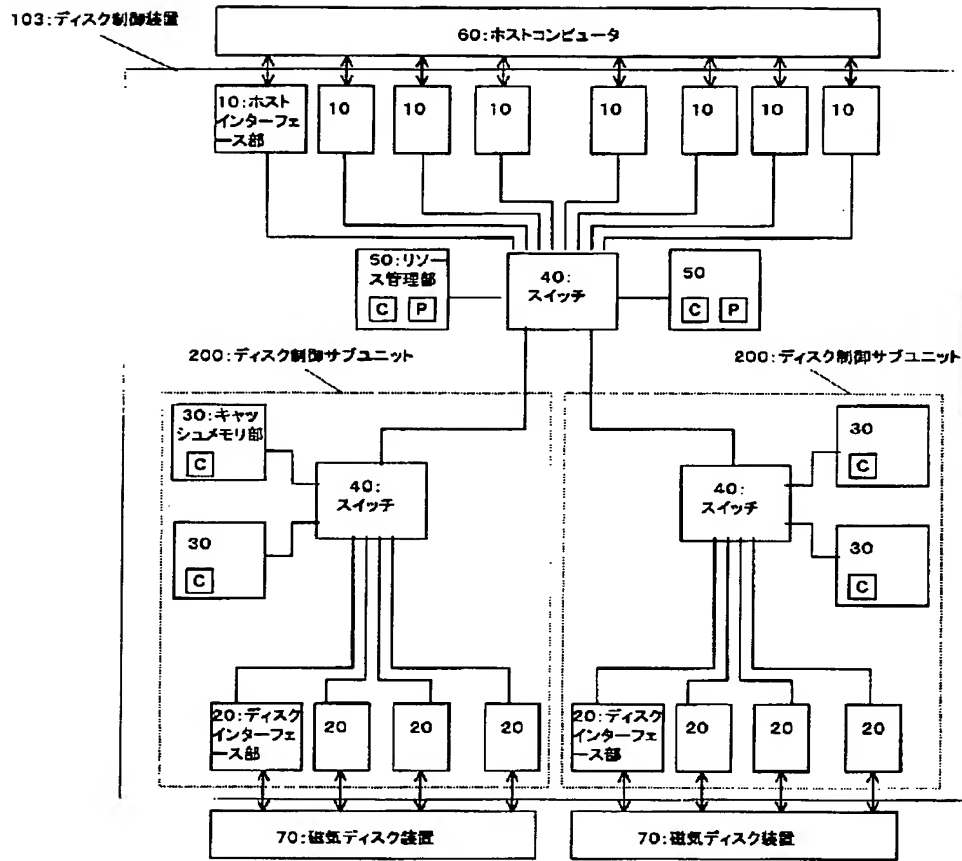
【図 4】

図 4



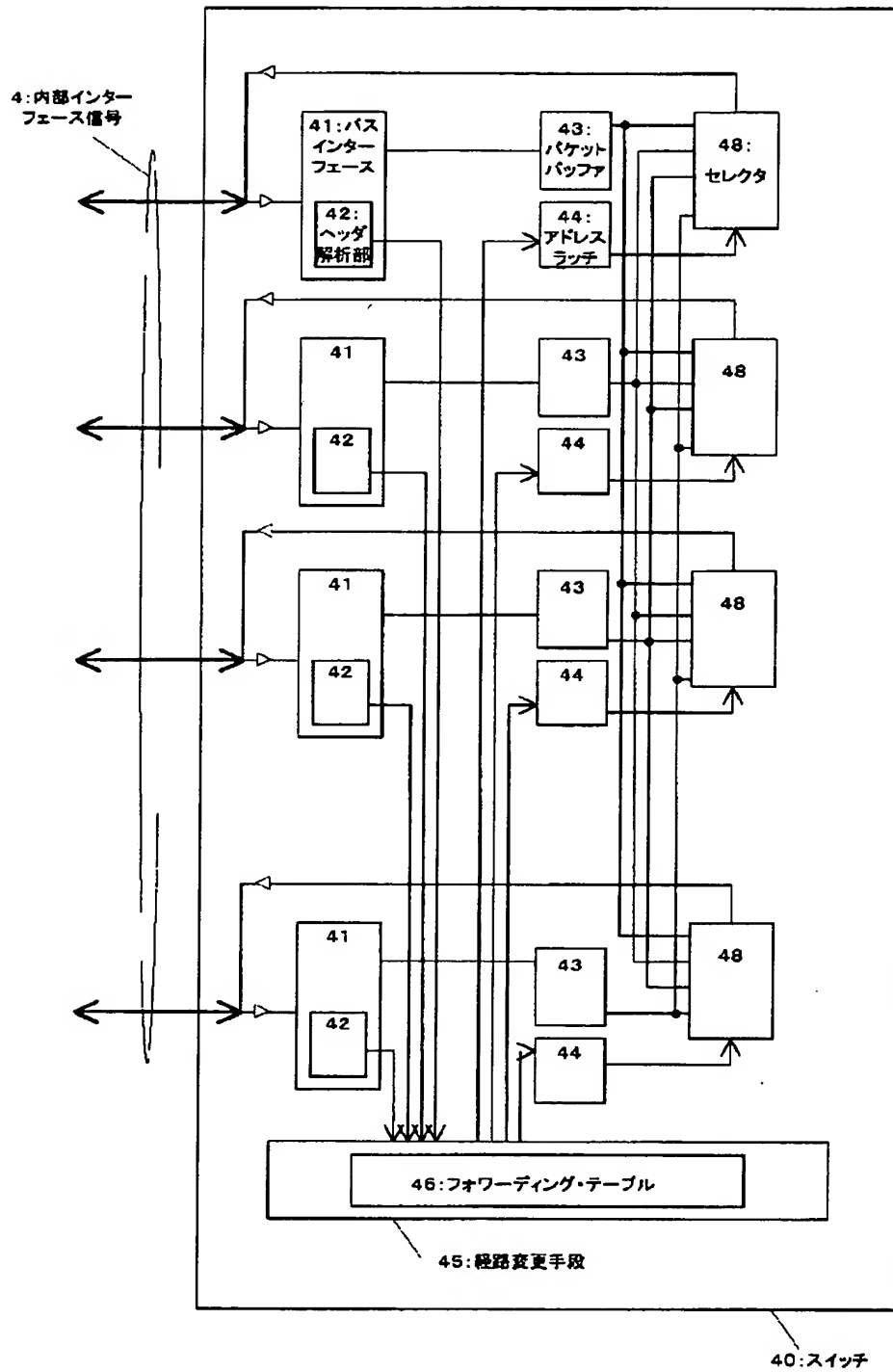
【図5】

図5



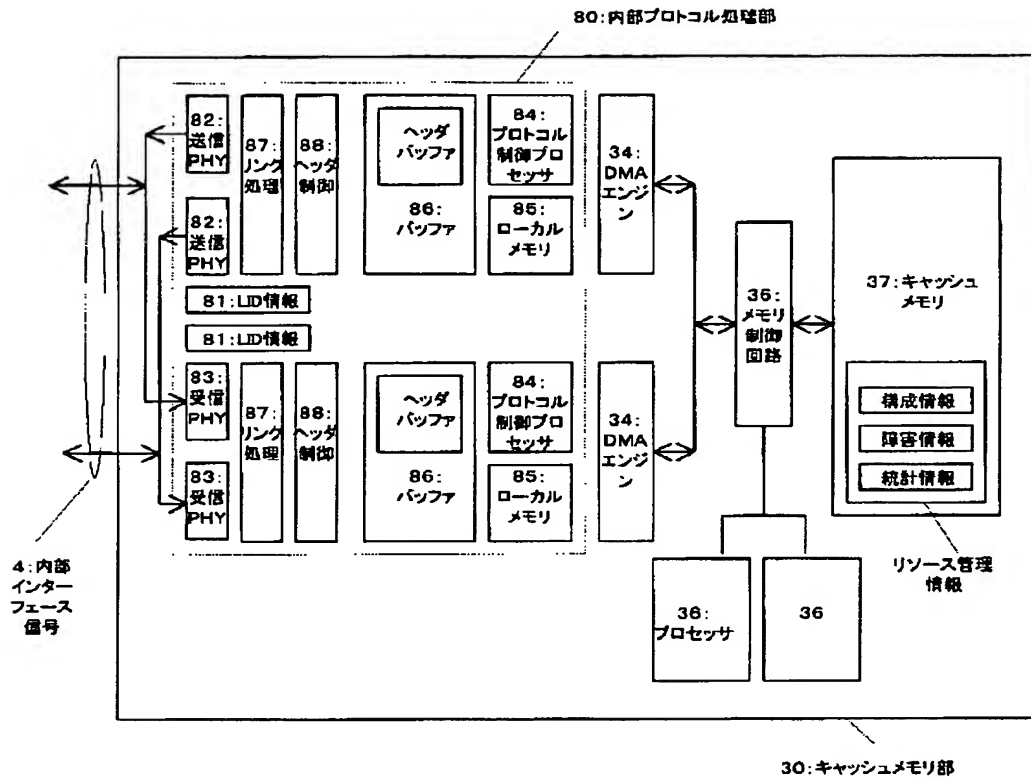
【図 6】

図6



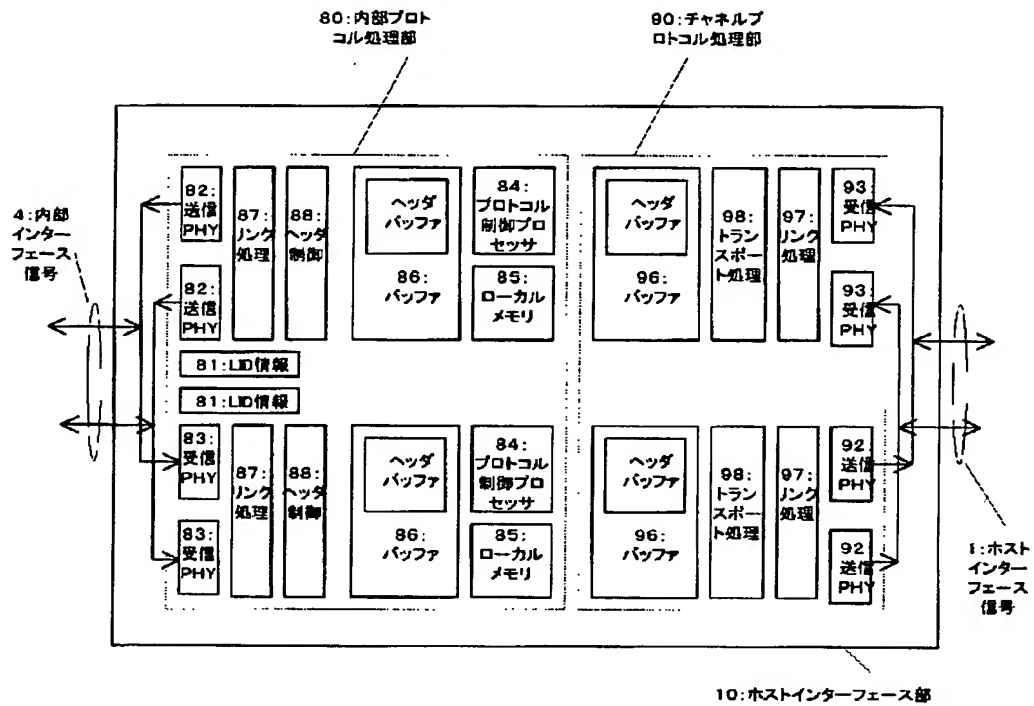
【図7】

図7



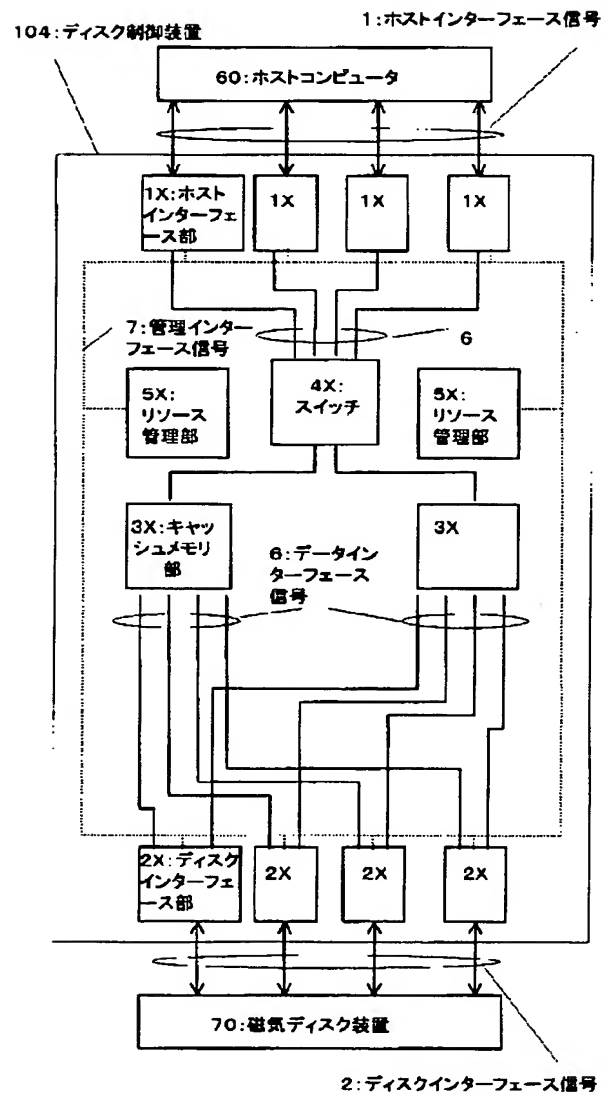
【図 8】

図8



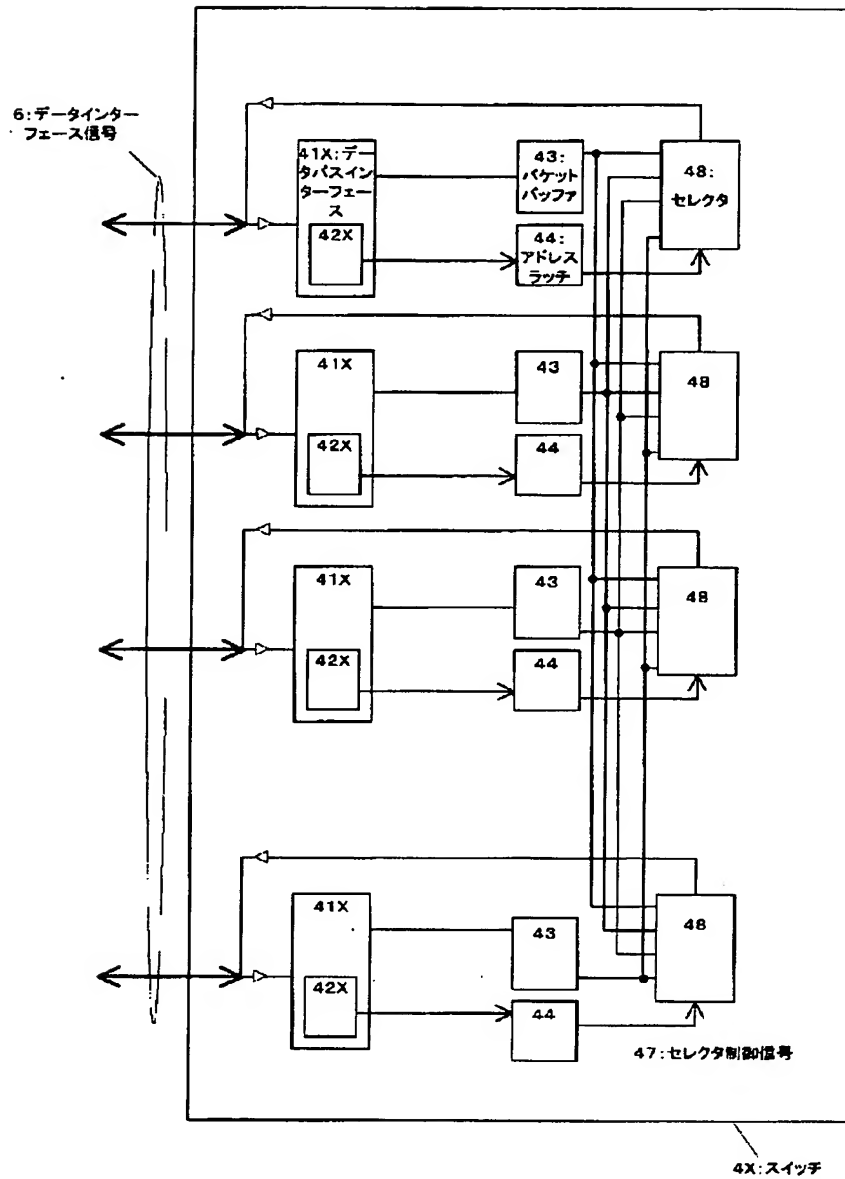
【図9】

図9



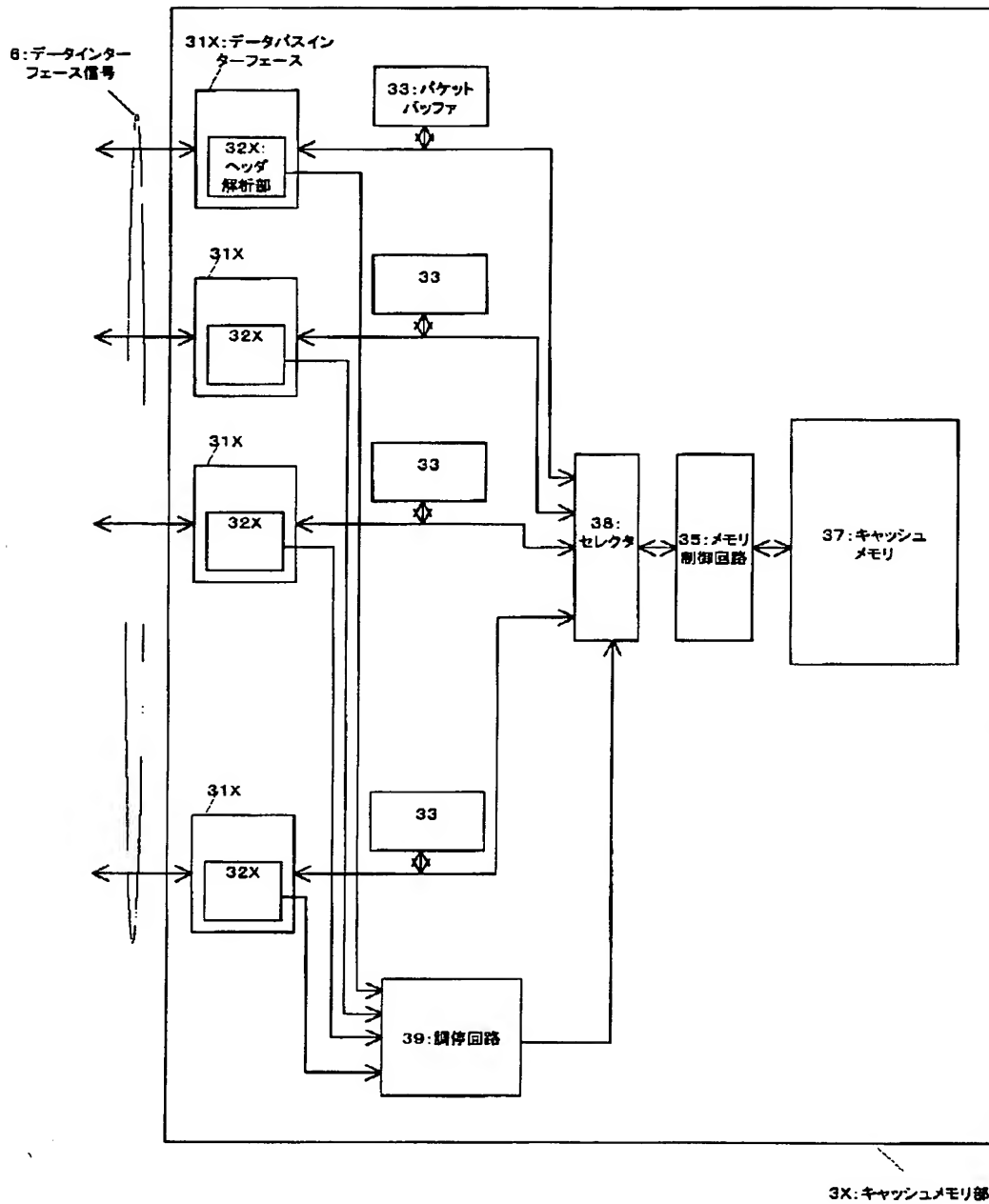
【図 1 0】

図10



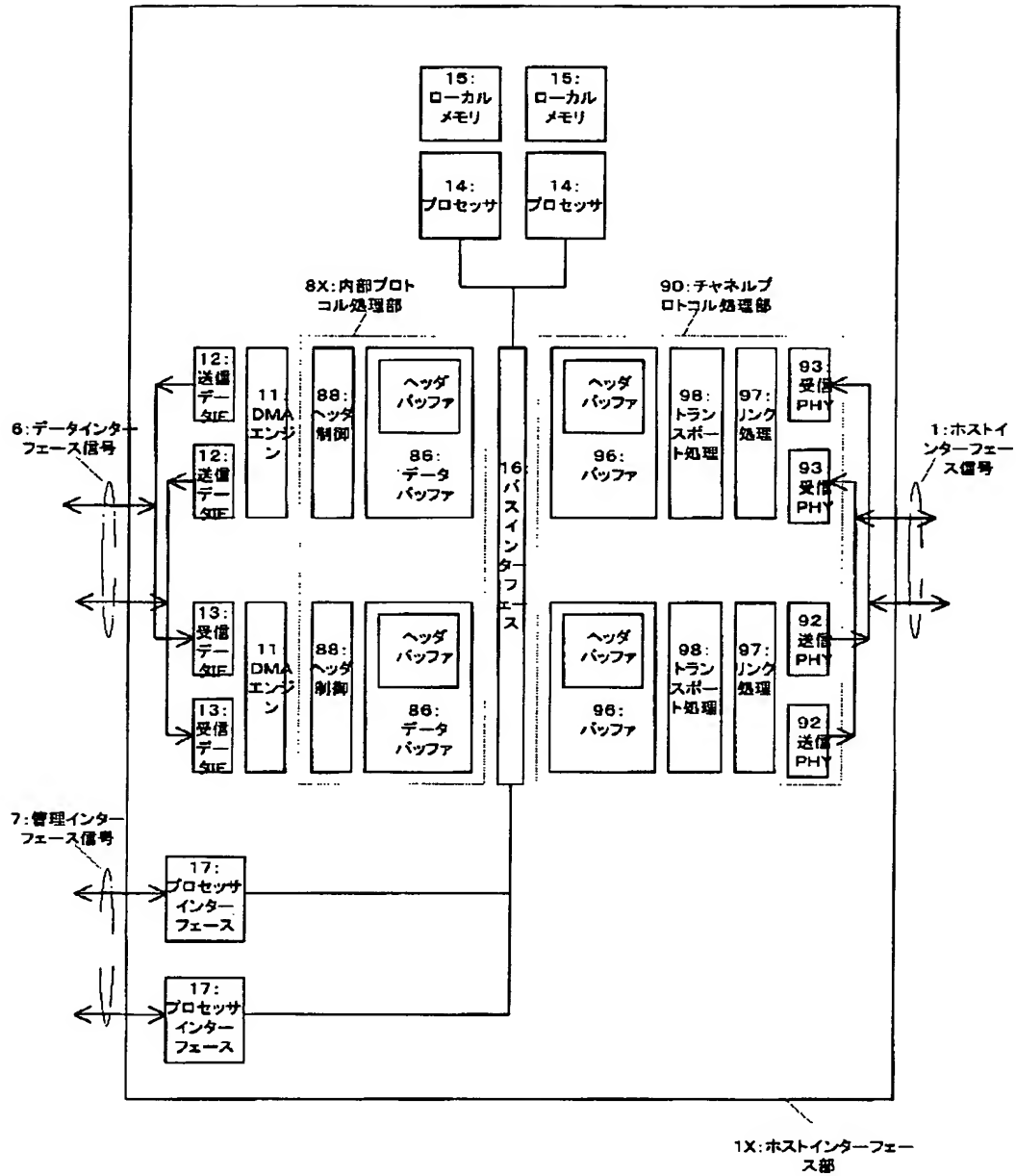
【図 11】

図11



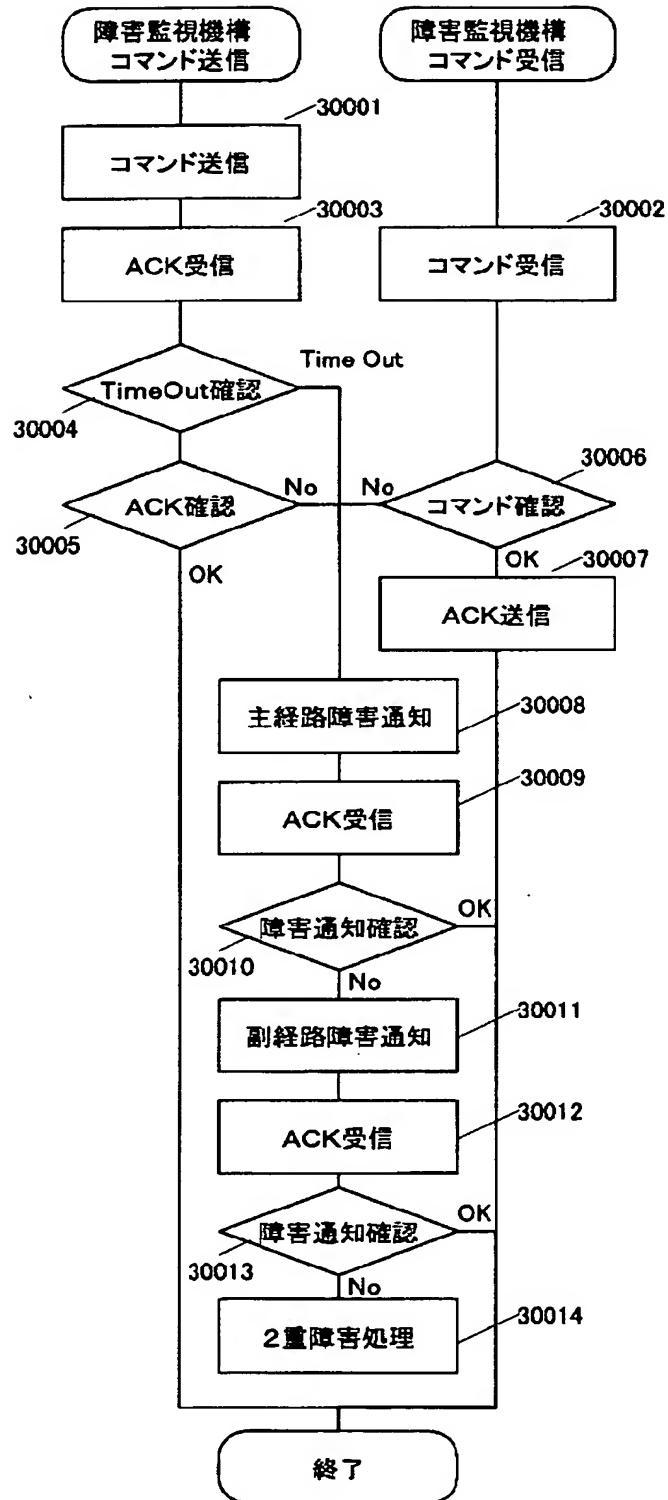
【図 12】

図12



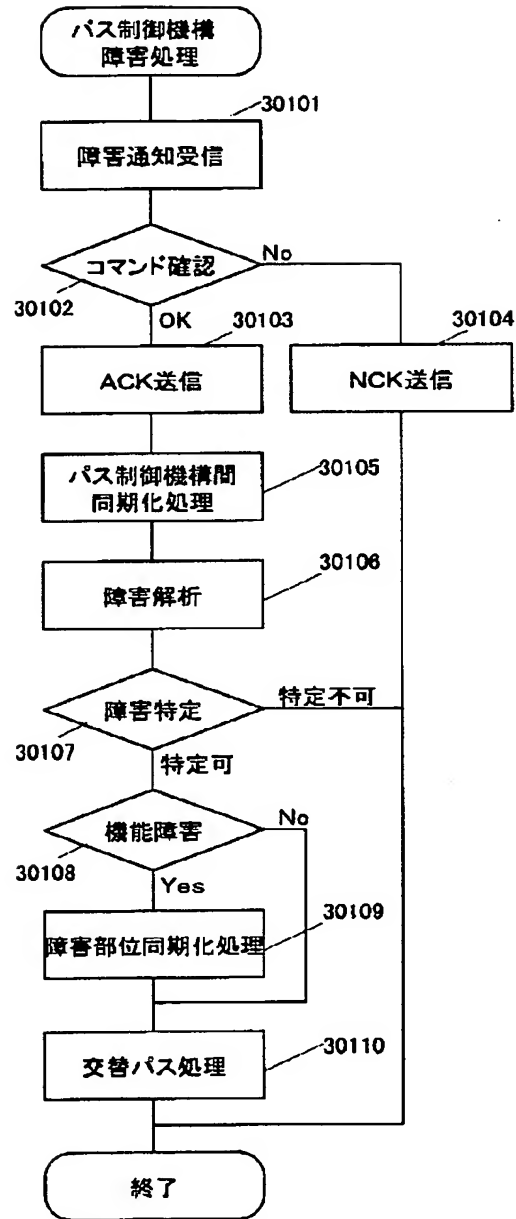
【図13】

図13



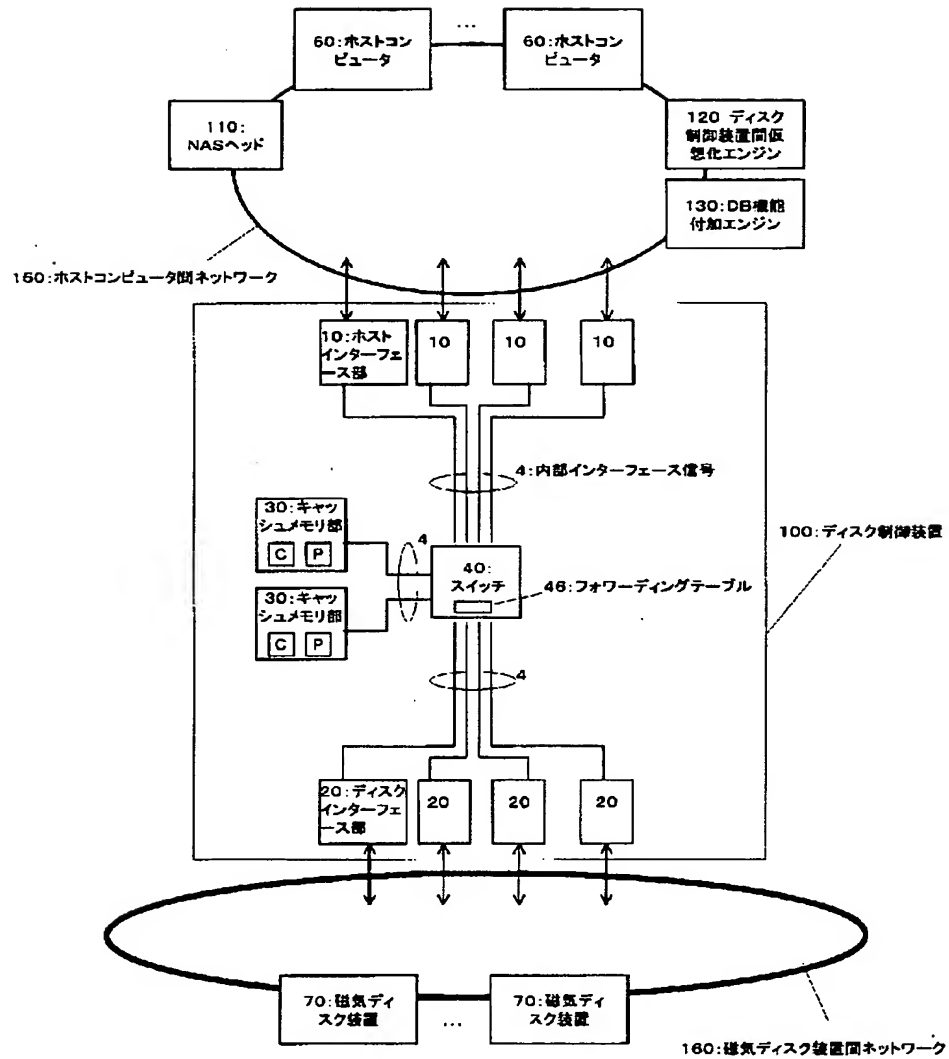
【図 1 4】

図14



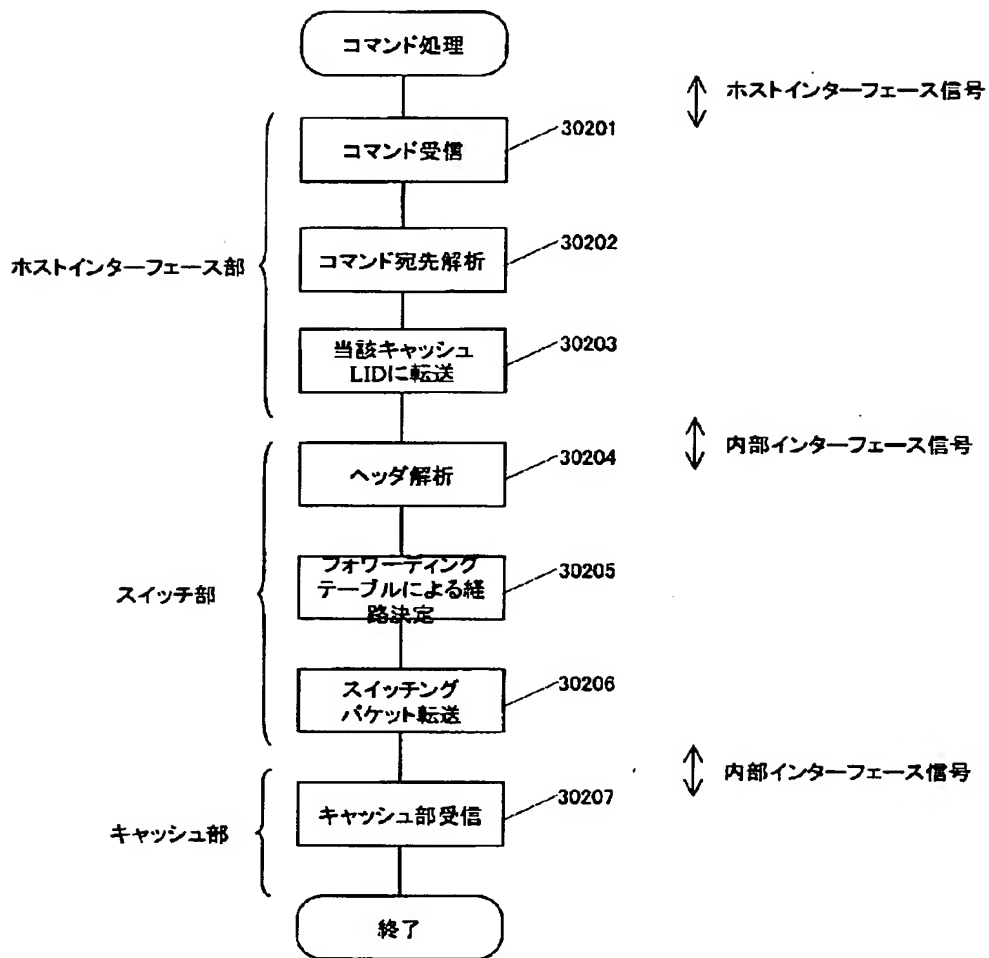
【図15】

図15



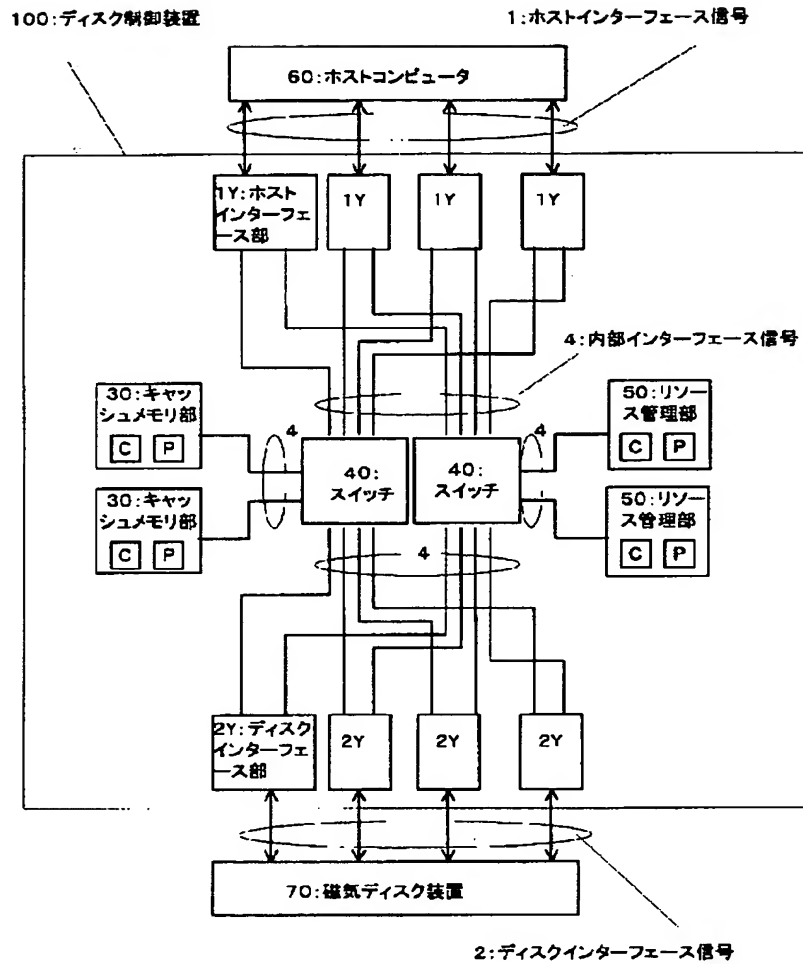
【図 1 6】

図16



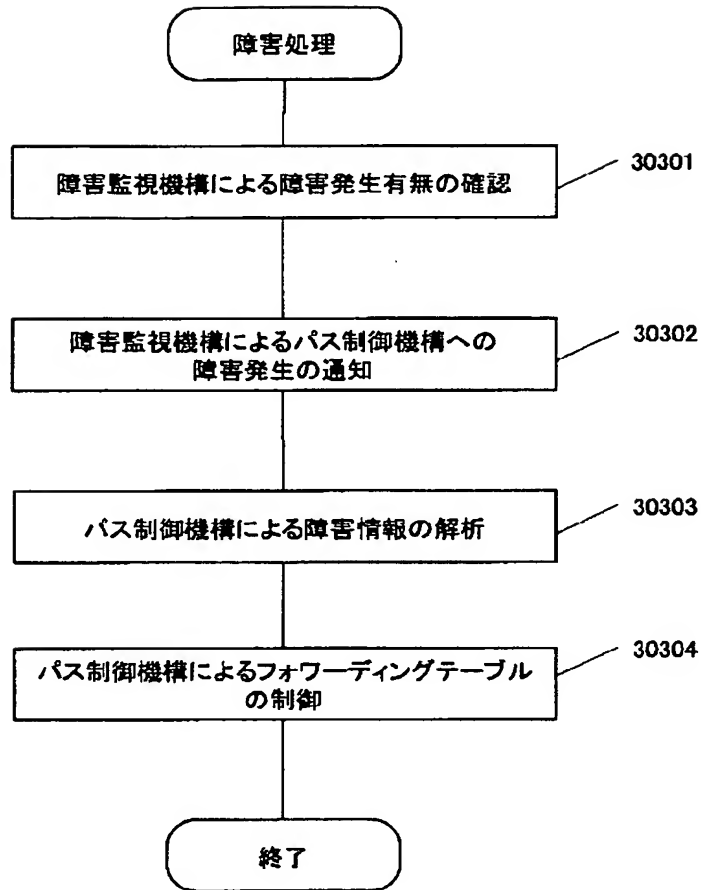
【図 17】

図17



【図 1 8】

図18



【書類名】                      要約書

【要約】

【課題】    ストレージ・システムの性能劣化や、ホスト・アプリケーションの動作不良を引き起こすことのない高可用性ディスク制御装置を提供する。

【解決手段】    複数のホストインターフェース部及び複数のディスクインターフェース部とキャッシュメモリ部との間を、1つ以上のスイッチで構成されるスイッチ網を介して接続し、スイッチにスイッチ網内での経路を指定するためのフォーワーディングテーブルと該テーブルを変更する変更手段を設け、また、ホストインターフェース部、ディスクインターフェース部及びキャッシュメモリ部に、スイッチ網内で一意に決まるローカルIDと該IDを変更する変更手段を設け、さらに複数のキャッシュメモリ部に、その障害発生の有無を監視するための障害監視機構と障害発生時に障害部位を回避するように前記スイッチ内のフォーワーディングテーブルを制御するためのパス制御機構を設ける。

【選択図】    図 1

認 定 ・ 付 加 情 報

特許出願の番号	特願 2 0 0 2 - 3 7 8 9 5 6
受付番号	5 0 2 0 1 9 8 1 8 0 1
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 1 月 6 日

< 認定情報・付加情報 >

【提出日】 平成14年12月27日



出 願 人 履 歴 情 報

識別番号 [ 0 0 0 0 0 5 1 0 8 ]

1. 変更年月日 1 9 9 0 年 8 月 3 1 日  
[変更理由] 新規登録  
住 所 東京都千代田区神田駿河台 4 丁目 6 番地  
氏 名 株式会社日立製作所